Prediction of 2024 US election ...*

Colin Sihan Yang Lexun Yu Siddharth Gowda

October 30, 2024

We forecast the winner of the 2024 US presidential election using "poll-of-polls" by building a linear model.

1 Introduction

Election result forecasting has become an essential tool for analysts in political science and the public to predict the outcome of democratic process, such as the presidential election in the United States. Traditionally, individual polls have been used as a snapshot of voter sentiment, but they only reflect temporary changes in the performance of contestants, instead of a precise estimation of the election result. As discussed by Pasek (2015) and Blumenthal (2014), the aggregation of multiple polls, or "poll-of-polls," has become a popular technique to reduce individual survey errors and provide more accurate election forecasts. However, the traditional poll aggregation does not reflect dynamics of an election, especially with real-time changes and the introduction of new data. This creates a gap for a more adaptable model to predict the election result based on both polling data and additional variables, such as historical data and economic indicators.

This paper fills the gap by building a hybrid election forecasting model following the strategies mentioned by Pasek (2015). As Pasek (2015) described in their article, aggregation involves determining which surveys are worth including, as well as selecting, combining and averaging results from multiple polls to reduce individual biases and errors. Prediction modeling adds other data to the model that predicts election outcomes based on current dynamics. Hybrid models like the Bayesian approach incorporates prior beliefs based on historical data or expert knowledge and new evidence like economic updates to dynamically adjust the forecast as the campaign progresses.

In this paper, we aim to predict the 2024 us election result with the hybrid election forcasting model. We incorporate aggregation by filtering the polls on FiveThirtyEight (2024) by

*Code and data are available at: https://github.com/yulexun/uselection.

numeric grade that indicates pollster's reliability, prediction that incorporates social and economic indicators including unemployment rates and abortion rates, and hybrid approaches that leverages Bayesian techniques which combines historical data such as the 2016 election data, allowing for a dynamic prediction of the U.S. presidential election.

The estimand for this research paper is the predicted support percentages for Kamala Harris and Donald Trump. The prediction is based on quantifying various polling factors, including sample size, poll scores, and transparency scores, which are used as predictors.

The results of this model indicate a more stable and accurate forecast compared to traditional aggregation methods alone, [update this …]

The remainder of this paper is structured as follows: [update this …]

2 Data

2.1 Overview

For the data we used in this analysis about the polling result for Kamala Harris and Donalad Trump in 2024 USA president election.
- **response variable:** pct(pct: The percentage of the vote or support that the candidate received in the poll)
- **numeric predictor:**
sample size(sample_size: The total number of respondents participating in the poll)
timegap(the time gap between the poll start date and the real election date i.e timegap = real US election date - poll start date)
pollscore(A numeric value representing the score or reliability of the pollster in question)
- **categorical predictor** state(The U.S. state where the poll was conducted or focused)
methodology(The method used to conduct the poll)

2.2 Measurement

In this dataset, each row represents a polling question that records the variables of interest. Each entry allows us to explore the real-world relationships between polling factors and the support percentage (`pct`) for the candidates Kamala Harris and Donald Trump. This dataset enables an analysis of how various polling characteristics influence the reported support levels for the candidates we are focused.

2.3 Clean Data

The data cleaning process involves several steps to ensure the quality and relevance of the polling data. First, we filter the dataset to retain only poll results with a numeric grade of 2.7 or higher, indicating that the polls are considered reliable. Next, we address missing values in the state attribute: polls with NA in the state column are considered national polls.

We then create a new attribute, days_taken_from_election, which represents the time gap between the poll's start date and the actual U.S. election date. Additionally, we filter the dataset to include only polls conducted after July 21, 2024, the date when Kamala Harris declared her candidacy. Finally, we remove any remaining rows that contain missing values to ensure a clean dataset.

Table 1: Sample of cleaned US election data

pct	sample_size	pollscore	days_taken_from_election	state	methodology	candidate_name
47.6	4180	-0.8	24	National	Online Ad	Kamala Harris
50.7	4180	-0.8	24	National	Online Ad	Donald Trump
0.8	4180	-0.8	24	National	Online Ad	Jill Stein
0.1	4180	-0.8	24	National	Online Ad	Chase Oliver
0.1	4180	-0.8	24	National	Online Ad	Cornel West
48.1	4180	-0.8	24	National	Online Ad	Kamala Harris

2.4 Basic Statistics Summary for Data

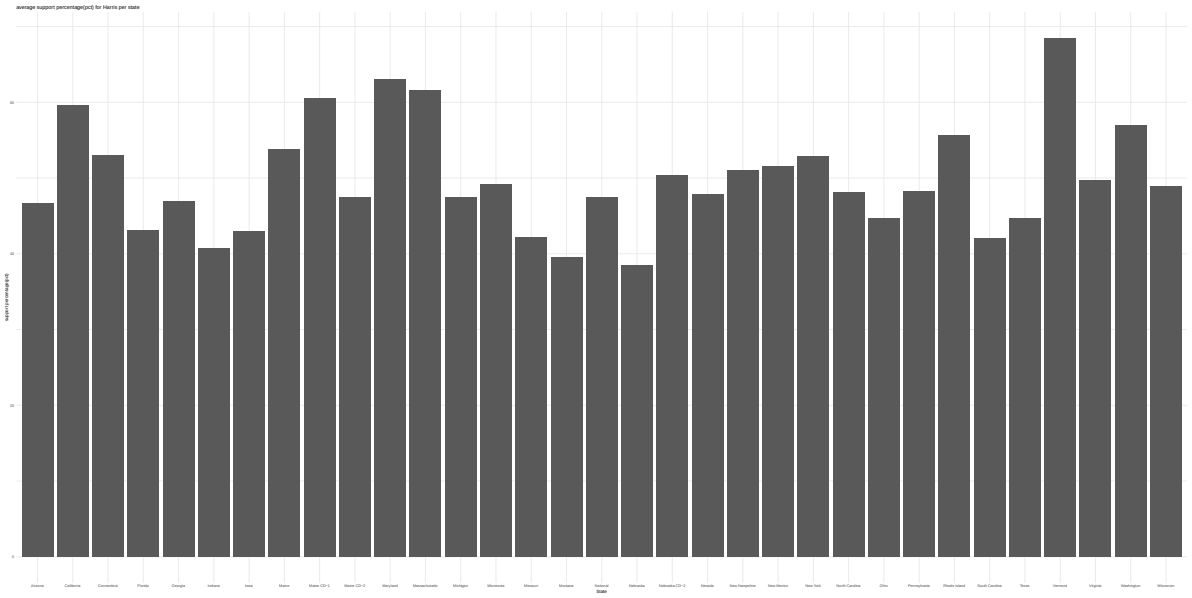
3 Model

The goal of our modelling strategy is twofold. Firstly,...

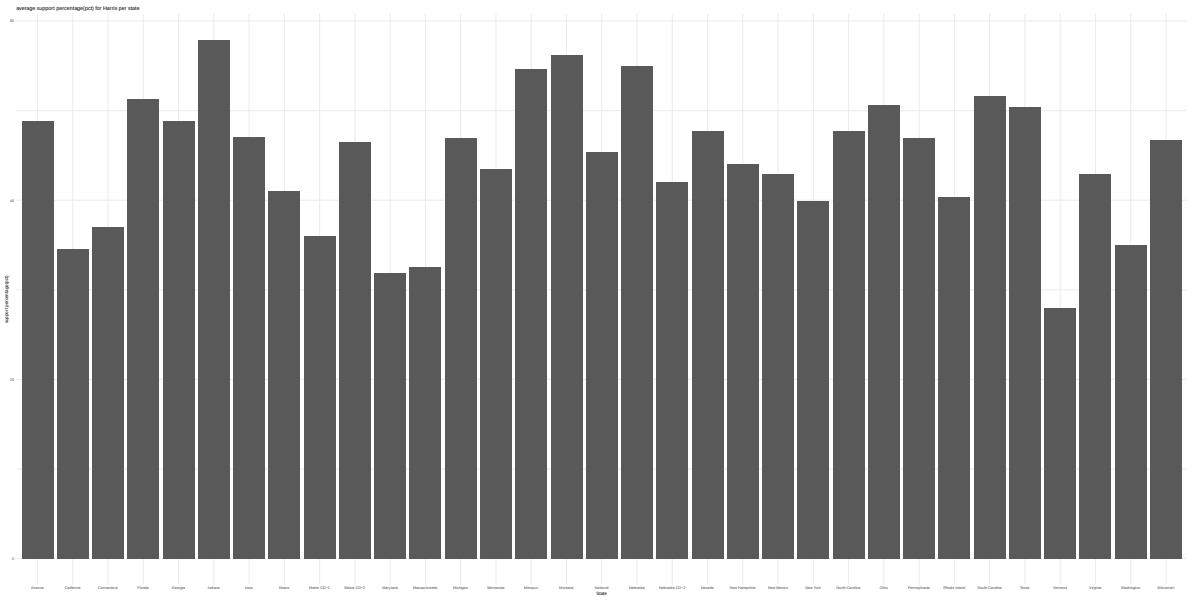
Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.



(a) average PCT vs State for Harris



(b) PCT vs State for Trump

Figure 1: the average PCT vs State for Harris and Trump

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Brilleman et al. (2018). We use the default priors from `rstanarm`. us

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

C Methodology of YouGov

YouGov's methodology documentations are seperated in two articles. The article by Bailey and Rivers (2024) documents the methodology of the 2024 election projection, while the webpage on YouGov (n.d.) documents the general methodology of YouGov's prediction.

C.1 Population, Frame, and Sample

As Bailey and Rivers (2024) stated, the population covered by YouGov's MRP model is everyone in the national voter file, whether or not they belong to YouGov's panel. The national voter files are digital database built by commercial organizations with public government records of voters, as explained by DeSilver (2018). Voter files indicates whether asomeone voted in a given election, thus YouGov's population covers all voters in previous US elections.

YouGov's sampling frame consists of its online panel members. These members are part of the SAY24 project, a collaboration between Stanford, Arizona State, and Yale Universities, as stated by Bailey and Rivers (2024). YouGov collect information on respondents when they join their panel before they are invited to participate in the survey.

YouGov select the sample from the sampling frame based on their ability to match characteristics of the population of interest. YouGov interviews nearly 100,000 people in the first set of estimates. For the second set of estimates, YouGov didn't just start over with a new sample. They took the initial data from August and September and updated it with responses from more than 20,000 additional registered voters who were re-interviewed in late September and early October.

C.2 Sample Recruitment

Panelists are recruited through various online channels, including advertisements and partnerships with websites (YouGov n.d.). They must provide demographic details upon joining, which helps in selecting representative samples for each survey. When respondents complete a survey, they are awarded points that can be exchanged for money.

C.3 Sampling Approach and Trade-offs

YouGov uses non-probability sampling due to the compensation, an approach where not every individual has an equal chance of selection (YouGov n.d.). This method allows quick and cost-effective data collection. However, as YouGov (n.d.) writes the panelists must have an internet connection to participate. YouGov state that there is 95% of us population with internet access, thus the sample may be less representative of certain hard-to-reach populations, such as individuals with very slow internet access or without internet access.

C.4 Non-response Handling

YouGov apply statistical weighting to adjust for the differences between the sample and target population. The weight is based on demographic characteristics such as age, gender, race and presidential vote (YouGov n.d.). Additionally, quality control measures exclude unreliable responses to improve data accuracy. The respondents are offered a small incentive to decrease the non-response and increase participation.

C.5 Strengths and Weaknesses of the Questionnaire

YouGov's surveys are conducted online, which is very efficient for the respondents, and responses are weighted to enhance representativeness. The pollster can recruit a large amount of panelists because of the online format. Combining with online tracking technologies, the metadata provided by ther panelists can be verified easily.

As a non-probability sample, it might miss certain demographic groups not covered by the online population. While weighting improves accuracy, it cannot fully substitute the randomization found in probability sampling. Additionaly, the categories in the surbey is oversimplified with bias. For instance, in the poll result published by YouGov, gender is divided into Male and Female. Race is divided into White, Black, Hispanic and Other. This indicates a lack of representation.

D FiveThirtyEight Licenses

FiveThirtyEight's data sets are used and modified by us under the Creative Commons Attribution 4.0 International License.

References

Bailey, Delia, and Douglas Rivers. 2024. "How YouGov's MRP Model Works for the 2024 U.S. Presidential and Congressional Elections." *YouGov*. https://today.yougov.com/politics/articles/50587-how-yougov-mrp-model-works-2024-presidential-congressional-elections-polling-methodology.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science and Politics* 47 (2): 297–300. http://www.jstor.org/stable/43284537.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

DeSilver, Drew. 2018. "Q&A: The Growing Use of 'Voter Files' in Studying the U.S. Electorate." *Pew Research Center*. https://www.pewresearch.org/short-reads/2018/02/15/voter-files-study-qa/.

FiveThirtyEight. 2024. "Our Data." *FiveThirtyEight*. https://data.fivethirtyeight.com.

Pasek, Josh. 2015. "THE POLLS–REVIEW: PREDICTING ELECTIONS: CONSIDERING TOOLS TO POOL THE POLLS." *The Public Opinion Quarterly* 79 (2): 594–619. http://www.jstor.org/stable/24546379.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

YouGov. n.d. "Methodology." Accessed October 31, 2024. https://today.yougov.com/about/panel-methodology.